DISTRIBUTIONALLY ROBUST CLASSIFICATION FOR MULTI-SOURCE UNSUPERVISED DOMAIN ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised domain adaptation (UDA) is a statistical learning problem when the distribution of training (source) data is different from that of test (target) data. In this setting, one has access to labeled data only from the source domain and unlabeled data from the target domain. The central objective is to leverage the source data and the unlabeled target data to build models that generalize to the target domain. Despite its potential, existing UDA approaches often struggle in practice, particularly in scenarios where the target domain offers only limited unlabeled data or spurious correlations dominate the source domain. To address these challenges, we propose a novel distributionally robust learning framework that models uncertainty in both the covariate distribution and the conditional label distribution. Our approach is motivated by the multi-source domain adaptation setting but is also directly applicable to the single-source scenario, making it versatile in practice. We develop an efficient learning algorithm that can be seamlessly integrated with existing UDA methods. Extensive experiments under various distribution shift scenarios show that our method consistently outperforms strong baselines, especially when target data are extremely scarce.

1 INTRODUCTION

Many supervised learning algorithms rely on the assumption that the training and test data are drawn from the same underlying distribution. However, this assumption is often violated in real-world applications due to various factors such as environmental changes, sampling biases, or temporal dynamics, leading to distribution shifts between training and deployment environments (Beery et al., 2018; Zech et al., 2018; Volk et al., 2019). Such mismatches can result in significant performance degradation, posing a fundamental challenge to achieving reliable generalization in practice (Gulrajani & Lopez-Paz, 2020; Koh et al., 2021; Sagawa et al., 2021).

Among various challenges posed by distribution shifts, we focus on unsupervised domain adaptation (UDA; Ganin et al. (2016); Long et al. (2015; 2018); Ben-David et al. (2010)). In UDA, we are given labeled data from a source domain and unlabeled data from a target domain, where the distributions of the two domains are related but not necessarily identical. The discrepancy may arise from differences in the marginal distribution of the input (Shimodaira, 2000; Sugiyama et al., 2007), the conditional distribution of the output given the input (Jin et al., 2023; Cai et al., 2023), or both (Tachet des Combes et al., 2020; Wu et al., 2025). The central objective is to build a predictive model that generalizes to the target domain by effectively transferring knowledge from the source domain.

Two representative approaches in UDA are distribution alignment and pseudo-labeling methods. Distribution alignment reduces the discrepancy between source and target domains, either in the data space (Courty et al., 2017b) or in the feature space (Ganin et al., 2016; Long et al., 2018). In contrast, pseudo-labeling methods refine target predictions using a source-trained model (Amini & Gallinari, 2003; Kumar et al., 2020; Wei et al., 2020), inspired by early work on self-training (Lee, 2013). Despite their popularity, both distribution alignment and pseudo-labeling methods often face practical limitations. Two common scenarios are when the amount of unlabeled target data is limited, or when the source data contains spurious correlations. When target data are scarce, alignment estimates can become unreliable and pseudo-labels may be noisy, which can lead to poor

generalization (Liang et al., 2020; Xie et al., 2020). When spurious correlations are present—for example, when irrelevant features like background or color are strongly associated with labels—models often rely on these shortcuts, which do not transfer to the target domain and ultimately degrade performance (Arjovsky et al., 2019; Zhao et al., 2019; Liu et al., 2021).

In this paper, we propose a new method for UDA, with a particular focus on classification tasks, based on the framework of distributionally robust learning (also known as distributionally robust optimization; DRO[1]). Our approach is inspired by the maximin effect estimation method developed for regression problems with multi-source data (Meinshausen & Bühlmann, 2015). While originally motivated by the multi-source setting, our method can also be applied in single-source scenarios by simulating multiple pseudo-sources through random indexing.

The proposed method estimates the conditional distribution of the output given the input for each source domain and constructs an ambiguity set consisting of their mixtures. The mixing weights are allowed to vary, enabling the model to adjust how much it trusts each source, while the input distribution is permitted to shift within a small Wasserstein ball around the observed target inputs. By explicitly modeling both sources of uncertainty—(i) which conditionals to rely on and (ii) how the target inputs may vary—our method improves target prediction performance, particularly when target data are scarce or when source domains contain spurious correlations.

We empirically evaluate our approach on widely used benchmarks, including digit classification tasks (MNIST, SVHN, USPS) and spurious correlation datasets (Waterbirds, CelebA, Colored MNIST). Our experiments focus on two key scenarios: (1) standard UDA with limited unlabeled target data and (2) settings where spurious correlations hinder generalization. In both cases, our method substantially outperforms existing approaches in domain adaptation and robust learning.

Our main contributions can be summarized as follows:

- We propose a novel distributionally robust framework that jointly models uncertainty in the target covariate distribution and the conditional label distribution defined in the feature space.
- We develop a tractable minimax optimization algorithm for the proposed method, which can be seamlessly combined with existing UDA methods.
- We provide extensive experiments demonstrating substantial improvements over well-known baselines in two challenging scenarios: target data scarcity and spurious correlations.

In the following, we review key lines of related work, including modern approaches to UDA, DRO-based methods, and other robust methods addressing spurious correlations.

1.1 Related works

Unsupervised domain adaptation UDA has been applied in a wide range of areas, including computer vision (Hoffman et al., 2018), natural language processing (Gururangan et al., 2020), and healthcare (Kamnitsas et al., 2017). Popular approaches in UDA learn domain-invariant features by aligning marginal or class-conditional distributions (Long et al., 2015; Ganin et al., 2016; Sun & Saenko, 2016; Tzeng et al., 2017; Long et al., 2018), or by separating normalization statistics across domains (Chang et al., 2019). They also regularize decision boundaries (Shu et al., 2018) and exploit pseudo-labels—including source-free variants (Liang et al., 2020; Yang et al., 2021). Optimal-transport methods align either marginal input distributions (Courty et al., 2017b) or joint distributions (Courty et al., 2017a; Damodaran et al., 2018). Recent work explores concept-based interpretability, diffusion-based alignment, information-theoretic regularization, one-shot settings, and heterogeneous-modal adaptation (Peng et al., 2024; Xu et al., 2025; Tan et al., 2025; Zhang et al., 2025; Yang et al., 2025). However, these methods can fail in the presence of spurious correlations.

To address this limitation, recent studies attempt to disentangle causal and spurious features using unlabeled target data. Examples include counterfactual consistency, generative interventions, and self-training refinements (Chen et al., 2020; Yue et al., 2023; Wei et al., 2025). Similar approaches have also been explored in multi-source settings (Li et al., 2023; Liu et al., 2025). However, these

[1]While DRO was originally studied in the context of optimization, in this paper we use the term more broadly to refer to distributionally robust learning methods under distribution shift.

methods often rely on strong structural assumptions or are highly sensitive to the quality of initial pseudo-labels, making them fragile under scarce or imbalanced target data.

Distributionally robust optimization Distributionally robust framework has recently gained attention in various areas, including UDA. One line of work in DRO defines ambiguity sets via Wasserstein balls or f-divergences, which provide robustness to small perturbations of the source distribution (Ben-Tal et al., 2013; Duchi et al., 2021; Awad & Atia, 2023). Another line considers maximin-effect or GroupDRO formulations that address mixture shifts, where the target distribution differs in mixing proportions across subpopulations (Meinshausen & Bühlmann, 2015; Zhan et al., 2024; Wang et al., 2025; Guo, 2024; Sagawa et al., 2019). Maximin-effect approaches are closely related to our setting but have been mostly studied in regression tasks. We provide further discussion on this point in Section 3. Mixture-shift formulations, particularly GroupDRO, are also relevant when spurious correlations arise from changing group proportions. However, GroupDRO typically assumes access to group labels and does not explicitly use unlabeled target data for improving target prediction performance. Recently (Guo et al., 2025) studies DRO for classification, but this paper mainly focuses on statistical convergence rates and the construction of bias-corrected estimators.

Other robust methods under spurious correlation A closely related setting to UDA is domain generalization (DG), where unlike UDA, no unlabeled target data are available during training. Representative DG methods, such as IRM (Arjovsky et al., 2019), V-REx (Krueger et al., 2021), and Fish (Shi et al., 2021), aim to learn representations that are invariant across multiple training environments to mitigate spurious correlations. Beyond domain generalization, other lines of work address spurious correlations more directly. These include approaches such as sample reweighting, biased auxiliary models, and adversarial weighting without group labels (Nam et al., 2020; Lahoti et al., 2020; Sohoni et al., 2020; Liu et al., 2021). Additional strategies suppress spurious features through contrastive objectives or mixup-based regularization (Zhang et al., 2022; Jin et al., 2024).

2 PRELIMINARIES

This section introduces basic notations, definitions and some preliminary setups. Let (X, Y) be the random variables of the input and output pair, where $X \in \mathcal{X} \subseteq \mathbb{R}^d$ and $Y \in \mathcal{Y}$. For a joint distribution P of (X, Y), its marginal distribution of X and conditional distribution of Y given X are denoted by P_X and $P_{Y|X}$, respectively. Also, we use the lower case p for denoting the density of the corresponding distribution denoted by the upper case P, and vice versa. The expectation with respect to the distribution P is denoted by $\mathbb{E}_P$.

Suppose for a moment that there is a single source domain. Let the population distributions of the source and target domains be denoted by P^{sc} and P^{tg}, respectively. For a function f^θ parameterized by $\theta \in \Theta \subseteq \mathbb{R}^p$, let $\ell(f^\theta(X), Y)$ be the loss function, which is the cross-entropy loss in most cases. In UDA, the goal is to minimize the target risk $\mathbb{E}_{P^{\mathrm{tg}}}[\ell(f^\theta(X), Y)]$ with the labeled source and unlabeled target data. The empirical risk minimizer (ERM) trained on the source data is the most naive baseline in UDA. However, when P^{sc} and P^{tg} are different, ERM can exhibit poor predictive performance on the target domain.

An important alternative to ERM estimator under distributional shift is DRO. Let $\mathcal{Q}$ denote a class of joint distributions over (X, Y), often referred to as the ambiguity set. In general, the DRO estimator with ambiguity set $\mathcal{Q}$ can be obtained by solving the minimax optimization problem

$$\underset{\theta \in \Theta}{\text{minimize}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[\ell(f^\theta(X), Y)]. \tag{1}$$

Constructing a suitable ambiguity set $\mathcal{Q}$ is a crucial and central component of DRO methods. A common approach is to define $\mathcal{Q}$ as a small (pseudo-)metric ball around the empirical distribution of source data or one of its variants. In practice, a key consideration is that the inner supremum must be not only well-defined but also computationally tractable, as it directly affects the overall optimization. For this reason, popular choices for the underlying metric include the Wasserstein distance (Gao et al., 2024; Gao & Kleywegt, 2023; Mohajerin Esfahani & Kuhn, 2018; Blanchet et al., 2021) and f-divergences (Duchi & Namkoong, 2021; Namkoong & Duchi, 2016), which allow for efficient approximations or closed-form solutions in many settings.

While ambiguity sets around the source distribution provide robustness to minor shifts, they are less suitable for structured shifts. A notable case is subpopulation shift, where the source is a mixture of

subpopulations and the target has the same subpopulations but different mixing weights. Here, the target can be far from the source in standard distances, making small-ball sets inadequate. A crucial example of an alternative formulation is the ambiguity set used in GroupDRO (Sagawa et al., 2019), which is specifically designed to address subpopulation shifts.

3 PROPOSED DISTRIBUTIONALLY ROBUST LEARNING METHOD

In this section, we present our multi-source DRO framework together with the corresponding learning algorithm. We assume access to labeled multi-source datasets $\mathbf{D}^{(k)} = \{(x_i^{(k)}, y_i^{(k)})\}_{i=1}^{N^{(k)}}$, where $\mathbf{D}^{(k)}$ denotes the dataset from the kth source domain ($k = 1, \ldots, K$), and an unlabeled target dataset $\mathbf{D}^{\mathrm{tg}} = \{x_i^{\mathrm{tg}}\}_{i=1}^{N^{\mathrm{tg}}}$. Although the framework is motivated by the multi-source setting, it can also be applied to single-source problems, as discussed in Section 3.1. Section 3.2 provides a high-level formulation of the proposed ambiguity set, highlighting the core intuition behind our method. Section 3.3 specifies the components of the ambiguity set in greater detail, and Section 3.4 presents the detailed learning algorithm.

3.1 MULTI-SOURCE FRAMEWORK FOR SINGLE-SOURCE PROBLEMS

Suppose that the source dataset $\mathbf{D}^{\mathrm{sc}} = \{(x_i^{\mathrm{sc}}, y_i^{\mathrm{sc}})\}_{i=1}^{N^{\mathrm{sc}}}$ come from a single domain. If a natural grouping variable is available, such as an environment or a demographic attribute, it can be used to partition $\mathbf{D}^{\mathrm{sc}}$ into multiple sources. Otherwise, each dataset $\mathbf{D}^{(k)}$ can be formed by random sub-sampling with replacement from $\mathbf{D}^{\mathrm{sc}}$. This sub-sampling procedure can be justified when the source distribution is a mixture of heterogeneous subpopulations: repeated random sub-sampling with replacement increases the chance that some sub-samples approximate an individual subpopulation. Consequently, when the target distribution consists of the same subpopulations but with different mixing weights, treating the sub-samples as distinct sources enables us to develop procedures that are robust to changes in mixture proportions. This idea has been used in the maximin effect (Meinshausen & Bühlmann, 2015), and our approach builds on the same principle. Throughout the paper, even in the single-source setting, we treat each $\mathbf{D}^{(k)}$ as if it were drawn from a distinct source domain, thereby framing the problem as a multi-source domain adaptation task. For notational and modeling convenience, we refer to these sub-samples as "sources."

3.2 PROPOSED AMBIGUITY SET: HIGH-LEVEL FORMULATION

Let $\hat{P}_{Y|X}^{(k)}$ denote the estimated conditional distribution based on the kth source $\mathbf{D}^{(k)}$, and let $\hat{P}_X^{\mathrm{tg}}$ be an estimator of the target input distribution. Details on how these estimators are constructed will be provided in Section 3.3. Given $\epsilon_1, \epsilon_2 \geq 0$ and $\bar{\beta} \in \Delta_{K-1}$, the proposed ambiguity set is defined as

$$\mathcal{Q} = \mathcal{Q}(\epsilon_1, \epsilon_2, \bar{\beta}) = \left\{ Q = (Q_X, Q_{Y|X}) \;\middle|\; \begin{array}{l} Q_{Y|X} = \sum_{k=1}^K \beta_k \hat{P}_{Y|X}^{(k)}, \quad \beta \in \Delta_{K-1} \\ D_1\left(Q_X, \hat{P}_X^{\mathrm{tg}}\right) \leq \epsilon_1, D_2\left(\beta, \bar{\beta}\right) \leq \epsilon_2 \end{array} \right\}, \tag{2}$$

where $\Delta_{K-1} = \{(\beta_1, \ldots, \beta_K) : \sum_{k=1}^K \beta_k = 1,\ \beta_k \geq 0\}$ denotes the $(K-1)$-dimensional simplex, and D_1 and D_2 are suitable divergence measures. Further details on the choices of D_1 and D_2 are provided in Section 3.3. The choices of $\hat{P}_{Y|X}^{(k)}$ and D_1 play a crucial role in ensuring the computational tractability of the proposed method.

The core idea behind the ambiguity set in (2) is that the target conditional distribution of the output given the input is expressed as a mixture of conditional distributions estimated from multiple source domains. To this end, we estimate different conditional distributions $\hat{P}_{Y|X}^{(k)}$ from sub-samples, and their mixtures form a sufficiently diverse class capable of containing the target conditional distribution. The radii ϵ_1 and ϵ_2 control the size of the ambiguity set and, consequently, the degree of distributional uncertainty allowed. The mixing vector β is centered around a reference vector $\bar{\beta}$, which can encode prior knowledge about the relative importance of each source. If no prior information is available, one can set $\bar{\beta}$ in proportion to the sample sizes of the source groups. In our experiments, each group was constructed with the same number of samples. Thus, we simply set $\bar{\beta}$ to the uniform vector.

In addition to modeling the conditional distribution of Y via mixtures, the ambiguity set (2) also accounts for uncertainty in the input distribution of the target domain through the radius parameter ϵ_1. This additional layer of robustness is particularly beneficial when the number N^{tg} of target samples is limited. In such cases, the empirical target input distribution $\hat{P}_X^{\mathrm{tg}}$ may be a poor estimate of the true P_X^{tg}, and small perturbations in the input space can lead to large variations in model performance. By allowing the input distribution to vary within a controlled neighborhood, the method hedges against this sampling variability and prevents overfitting to a small or biased target sample. Such robustness is especially useful in practice, as limited or imbalanced target samples often fail to capture the diversity of the true domain.

The proposed ambiguity set in (2) is inspired by recent work on maximin effect estimation in regression settings (Meinshausen & Bühlmann, 2015; Guo, 2024; Wang et al., 2025). These studies consider DRO formulations where the ambiguity set consists of mixtures of conditional distributions from multiple source domains. In contrast to our classification setting, their analysis is primarily on regression, using negative explained variance or squared error loss as the performance criterion. This choice offers computational benefits; for example, in linear regression it leads to a closed-form solution for the optimization problem (see Theorem 1 in Meinshausen & Bühlmann (2015)). In classification settings, however, developing a computationally feasible DRO algorithm under an ambiguity set of the form (2) presents new challenges. Addressing these challenges and designing an efficient algorithm for this setting constitutes one of our main technical contributions, which we describe in the following subsections.

3.3 Detailed specification of the ambiguity set

To fully specify the ambiguity set $\mathcal{Q}$ in (2), we define the estimators $\hat{P}_X^{\mathrm{tg}}$ and $\hat{P}_{Y|X}^{(k)}$ as well as the two divergence measures D_1 and D_2. The reference vector $\bar{\beta}$ is treated as fixed, while the radii ϵ_1 and ϵ_2 serve as tuning parameters that control the size of the ambiguity set.

We simply define $\hat{P}_X^{\mathrm{tg}}$ as the empirical measure based on $\mathbf{D}^{\mathrm{tg}}$. The construction of each $\hat{P}_{Y|X}^{(k)}$ depends on the available data. If sufficiently large datasets are available for each source, one can estimate $\hat{P}_{Y|X}^{(k)}$ separately using only the data from that source. However, in many practical settings, the amount of data per source may be limited. Moreover, when the distributions of different sources are similar, estimating them separately can be an inefficient use of data. To address this problem, in our experiments, we first train a single classification model using the entire source dataset. From this trained model, we extract a feature map $z : \mathcal{X} \to \mathcal{Z}$ by removing the final classification layer. Given the extracted features, we then estimate each $\hat{P}_{Y|X}^{(k)}$ (or the corresponding conditional density $\hat{p}_{Y|X}^{(k)}$) using a simple linear logistic regression trained independently on each source, treating the softmax outputs as practical probability estimates. This approach ensures that the proposed method remains efficient and scalable, even when the number of sources K is large.

It is worth noting that the proposed method can be combined with existing UDA methods. For instance, after obtaining the feature map z, one may employ well-known approaches such as CDAN or STAR to construct $\hat{P}_{Y|X}^{(k)}$ using z as the initial feature map, leading to further performance improvements. Further explanation and empirical results of this combination are provided in Section 4.2.

For the divergence D_2, we simply use the Euclidean distance. For D_1, we adopt the infinite-order Wasserstein distance, chosen for its computational tractability; see Section 3.4 for algorithmic details. Recall that the Wasserstein distance of order $t \in [1, \infty)$ between two probability measures P and Q is defined as $W_t(Q, P) = \inf_\gamma \{(\int c(x, x')^t \, d\gamma)^{1/t}\}$, where the infimum is taken over all couplings γ of Q and P, and $c(\cdot, \cdot)$ is a cost function. The infinite-order Wasserstein distance is then given by $W_\infty(Q, P) := \lim_{t\to\infty} W_t(Q, P)$.

In our case, we define the cost function as the Euclidean distance between feature representations, that is, $c(x, x') = \|z(x) - z(x')\|_2$, where z denotes the feature map introduced earlier. This choice is often more effective than defining the cost function directly in the raw input space, particularly in high-dimensional settings (Ganin et al., 2016; Zeiler & Fergus, 2014; Krizhevsky et al., 2017).

Algorithm 1 Procedure for minimizing (3)

1: **Input:** step sizes $\eta_\theta, \eta_\beta, \eta_z$; initial values $\theta^{(0)}$, $\beta^{(0)}$
2: **repeat** for $t = 1, 2, \ldots$
3: Sample $x_i^{\mathrm{tg}} \sim \hat{P}_X^{\mathrm{tg}}$ and compute $z_i^{\mathrm{tg}} = z(x_i^{\mathrm{tg}})$
4: Compute $y^\circ(\beta^{(t-1)}, x_i^{\mathrm{tg}})$ as in (4)
▷ Step 1: Update z'
5: Update z_i' using the gradient ascent and projection steps in (5)
▷ Step 2: Update β
6: Update $\beta^{(t)}$ using the exponentiated gradient ascent and projection steps in (6)
7: Compute $y^\circ(\beta^{(t)}, x_i^{\mathrm{tg}})$
▷ Step 3: Update θ
8: Update $\theta^{(t)}$ using the gradient descent in (7)
9: **until** convergence

3.4 Learning algorithm

The proposed DRO estimator is obtained by solving the general DRO problem (1) with the ambiguity set defined in (2). In this subsection, we describe the detailed learning algorithm. We assume that $f^\theta(X)$ depends on X only through the fixed representation $z(X)$ as defined in Section 3.3. Thus, we denote the model by $f_Z^\theta(z(X))$ in the remainder of this section. For notational convenience, we often treat the probability mass function $\hat{p}_{Y|X}^{(k)}(\cdot \mid X)$ as a probability vector. Before presenting the algorithm, we state the following proposition, which facilitates a computationally efficient implementation.

Proposition 3.1 (Tractable surrogate reformulation)**.** *Suppose that the feature map $z : \mathcal{X} \to \mathcal{Z}$ and the estimators $\hat{P}_{Y|X}^{(k)}$ are given. Let the divergence measures D_1 and D_2 be defined as in Section 3.3. For any $\epsilon_1, \epsilon_2 > 0$ and a fixed θ, the maximization objective in the DRO problem* (1) *with the ambiguity set $\mathcal{Q}$ defined in* (2)*, i.e.,* $\sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[\ell\left(f^\theta(X), Y\right)]$ *is upper-bounded by the following surrogate objective*

$$\sup_{\substack{\beta \in \Delta_{K-1}, \\ \|\beta - \bar{\beta}\|_2 \le \epsilon_2}} \mathbb{E}_{\hat{P}_X^{\mathrm{tg}}} \left[\sup_{\|z' - z(X)\|_2 \le \epsilon_1} \ell\left(f_Z^\theta(z'), y^\circ(\beta, X)\right) \right], \tag{3}$$

where

$$y^\circ(\beta, x) := \sum_{k=1}^{K} \beta_k \, \hat{p}_{Y|X}^{(k)}(\cdot \mid x) \quad \textit{for } x \in \mathcal{X} \tag{4}$$

is the soft pseudo-label vector defined as a convex combination of source conditionals.

This reformulation admits an efficiently computable algorithm that leverages adversarial feature perturbation z' and soft pseudo-label y°. The detailed rationale behind the approximation (3) is provided in Appendix A.1.

We now outline the procedure for minimizing (3). At a high level, the algorithm alternates between updating z', β, and θ using a minimax optimization scheme implemented via stochastic gradient descent, as summarized in Algorithm 1. The detailed update rules are described below.

1. Update of z'. Suppose that θ and β are fixed. Given a sample x_i^{tg} from $\hat{P}_X^{\mathrm{tg}}$, let $z_i^{\mathrm{tg}} = z(x_i^{\mathrm{tg}})$. To maximize the mapping $z' \mapsto \ell\left(f_Z^\theta(z'), y^\circ(\beta, x_i^{\mathrm{tg}})\right)$, we perform a projected gradient ascent starting from z_i^{tg}:

$$\begin{aligned} z_i' &\leftarrow z_i^{\mathrm{tg}} + \eta_z \nabla_z \ell\left(f_Z^\theta(z_i^{\mathrm{tg}}), y^\circ(\beta, x_i^{\mathrm{tg}})\right) \\ z_i' &\leftarrow \Pi_{\{z : \|z - z_i^{\mathrm{tg}}\|_2 \le \epsilon_1\}}(z_i'), \end{aligned} \tag{5}$$

where ∇_z denotes the gradient of $z \mapsto \ell(f_Z^\theta(z), y^\circ(\beta, x_i^{\mathrm{tg}}))$, η_z is the step size, and $\Pi_{\mathcal{A}}(\cdot)$ is the Euclidean projection onto the set $\mathcal{A}$.

2. Update of β. Given θ and z'_i, we update the mixture weights β via a projected exponentiated gradient ascent (Sagawa et al., 2019):

$$\beta_k \leftarrow \frac{\beta_k \exp\left(\eta_\beta \ell\left(f_Z^\theta(z'_i), \hat{p}_{Y|X}^{(k)}(\cdot \mid x_i^{\mathrm{tg}})\right)\right)}{\sum_{j=1}^K \beta_j \exp\left(\eta_\beta \ell\left(f_Z^\theta(z'_i), \hat{p}_{Y|X}^{(j)}(\cdot \mid x_i^{\mathrm{tg}})\right)\right)}, \quad \text{for } k = 1, \ldots, K,$$
$$\beta \leftarrow \Pi_{\{\beta : \|\beta - \bar{\beta}\|_2\} \le \epsilon_2}(\beta). \tag{6}$$

where η_β is the step size. See Appendix A.2 for further details on this update.

3. Update of θ. Given z'_i and β, we update θ using a stochastic gradient descent:

$$\theta \leftarrow \theta - \eta_\theta \nabla_\theta \ell\left(f_Z^\theta(z'_i), y^\circ(\beta, x_i^{\mathrm{tg}})\right), \tag{7}$$

where ∇_θ denotes the gradient of the map $\theta \mapsto \ell(f_Z^\theta(z'_i), y^\circ(\beta, x_i^{\mathrm{tg}}))$.

4 Experiments

In this section, we evaluate the proposed method in two experimental settings using widely adopted benchmark datasets. The first experiment considers digit recognition tasks across three domains—MNIST, SVHN, and USPS—and compares the proposed method with existing UDA approaches, with particular emphasis on scenarios where the amount of target-domain data is limited. The second experiment examines the robustness of our method in the presence of spurious correlations, using popular benchmarks such as Waterbirds, CelebA, and Colored MNIST, where distribution shifts are induced by non-causal attributes.

In the first experiment, one of MNIST, SVHN, or USPS is used as the single source domain, and one of the remaining two serves as the target domain, resulting in three source–target pairs. In the second experiment, each benchmark dataset consists of a single source domain and a single target domain as described below. Thus, all experiments are conducted under the single-source UDA setting. See Appendix A.3 for more detailed descriptions of the datasets.

MNIST, SVHN, and USPS (LeCun et al., 2002; Netzer et al., 2011; Hull, 2002): Three benchmark datasets for handwritten digit recognition from different domains. MNIST contains grayscale images of digits from 0 to 9, SVHN consists of color images of house numbers obtained from Google Street View, and USPS includes grayscale images of handwritten digits collected from postal envelopes. For each dataset, we randomly sampled 10^2 and 10 unlabeled target samples per class.

Waterbirds, CelebA, and CMNIST (Sagawa et al. (2019); Liu et al. (2015); Arjovsky et al. (2019)): Three benchmark datasets widely used to study spurious correlations. Each dataset is divided into four groups, where spurious attributes (e.g., background in Waterbirds, gender in CelebA, or color in CMNIST) are correlated with labels. In all cases, three majority groups are combined to form the single-source domain, and the remaining minority group serves as the target domain. For Waterbirds, CelebA, and CMNIST, the unlabeled target sample sizes N^{tg} are 56, 1,387, and 2,998, respectively.

4.1 Experimental setup

Following Meinshausen & Bühlmann (2015), in all our experiments we set $K = 10$ and draw each sub-sample of size $N = N^{\mathrm{sc}}/5$ with replacement, which we denote by $\mathbf{D}^{(k)} = \{(x_i^{(k)}, y_i^{(k)})\}_{i=1}^N$, $k = 1, \ldots, 10$. We then use the entire source dataset to learn the feature map z. For constructing a base classifier $\hat{P}_{Y|X}^{(k)}$, we use three approaches: training a simple linear logistic regression model (ERM) on $\{(z(x_i^{(k)}), y_i^{(k)})\}_{i=1}^N$, and training CDAN and STAR, both initialized with the learned feature map z. For the backbone, we use two different models. In the first setting, we adopt a deep neural network, following the architecture in (Ganin et al., 2016). In the second setting, we employ ResNet-50 following (Sagawa et al., 2019).

For hyperparameter selection, we follow a widely adopted procedure in domain adaptation benchmarks, which uses a small labeled target validation set for tuning. Specifically, we assume the

Method	SVHN→ MNIST		MNIST → USPS		USPS → MNIST	
	10^2	10	10^2	10	10^2	10
ERM (Src-only)	59.6± 1.8		63.4± 1.4		60.4± 5.7	
DANN	66.0± 4.9	61.2± 1.8	82.0± 3.9	74.3± 5.7	74.8± 6.7	51.1± 5.0
CDAN	63.4±1.8	56.9±0.9	80.8±1.4	62.0±1.8	58.3±4.6	54.8±5.9
MK-MMD	50.0±3.0	48.3±1.0	63.3±1.1	41.3±5.5	57.6±3.2	32.1±3.9
ATDOC	83.3±9.1	59.1±5.7	91.1±0.8	73.6±4.2	92.6±1.3	87.0±3.6
STAR	76.4±1.5	66.8±0.9	90.3±1.9	81.3±7.6	94.5±0.7	85.2±2.7
CORAL	75.4±2.7	63.6±0.9	90.4±0.7	85.4±0.5	75.9±1.9	64.7±3.5
MCD	79.1±1.0	61.3±0.7	89.3±1.6	84.5±2.3	96.1±1.6	85.9±4.0
Ours (ERM)	92.0±1.6	87.0±0.9	92.1±1.1	87.1±3.6	90.3±2.2	86.4±2.5
Ours (CDAN)	92.5±2.3	87.0±2.0	93.5±1.3	87.8±1.5	91.5±1.9	87.0±1.9
Ours (STAR)	**94.4**±1.7	**91.3**±1.1	**95.6**±1.0	**91.2**±1.4	**97.3**±0.8	**93.0**±2.8

Table 1: Comparison of test accuracies across different target sample sizes for three domain adaptation benchmarks: SVHN, MNIST, and USPS. Each block reports results under two target data sizes: 10^2 and 10 unlabeled samples per class. Here, Ours$(\cdot)$ denotes our methods built upon different base classifiers. Boldface indicates the best performance.

availability of a small validation set containing 10 labeled target samples per class. This set is used to perform a grid search over $\epsilon_1 \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ and $\epsilon_2 \in \{0, 0.2, 0.4, 0.6, 1\}$. While this approach is not strictly unsupervised, it is a practical compromise employed in prior works (Yue et al., 2023; Courty et al., 2017b; Saito et al., 2018). To ensure fairness, the same selection procedure is applied to all competing methods.

We compare our method with several baselines widely used in domain adaptation and robust learning: DANN (Ganin et al., 2016), CDAN (Long et al., 2018), MK-MMD (Long et al., 2015), CORAL (Sun & Saenko, 2016), MCD (Saito et al., 2018), ATDOC (Tang et al., 2020), STAR (Lu et al., 2020), GroupDRO, and ICON (Yue et al., 2023).

4.2 Results on digit recognition tasks

In the first experiment, we consider two target sample sizes: 10^2 and 10 per class. This setup allows us to evaluate whether our method can maintain robust performance even when only a very small number of target samples are available for training. Table 1 summarizes the results for three source–target domain pairs. Our method consistently achieves the best performance across all tasks and target sample sizes. Notably, the performance can be further improved by combining our approach with existing UDA methods such as CDAN and STAR. For example, when using CDAN as the base classifier, our method improves upon classical CDAN by $+29.1\%$ on the SVHN→MNIST task.

These improvements highlight the twofold strength of our framework. First, the ambiguity set explicitly accounts for uncertainty in the target input distribution through the ϵ_1-radius. This design is particularly beneficial under extreme data scarcity, where the empirical distribution $\hat{P}_X^{\mathrm{tg}}$ may be a poor approximation of the true P_X^{tg}. By allowing controlled perturbations, the model hedges against sampling variability and mitigates overfitting to small or biased target samples. Second, incorporating UDA methods such as CDAN or STAR enables the construction of conditional estimators that generalize better to the target domain. The formulation of the ambiguity set as a mixture of these refined conditionals produces an uncertainty set more closely aligned with the true target distribution. Consequently, the final model trained under this set achieves better generalization to target than mixtures based solely on ERM conditionals.

4.3 Results on spurious correlation benchmarks

Table 2 reports the test accuracies on the target domain across the three spurious-correlation benchmarks. Overall, most baseline methods show poor generalization to the target domain, with some performing even worse than standard ERM. This underscores the difficulty of these benchmarks, where spurious correlations and target-data scarcity occur simultaneously.

Method	Group label	Waterbirds Test Acc	CelebA Test Acc	CMNIST Test Acc
ERM (Src-only)	×	48.4±0.9	35.5±0.6	0.9±0.5
DANN	×	35.8±4.5	23.5±2.1	0.9±1.8
CDAN	×	46.2±1.8	24.6±1.5	1.2±0.4
MK-MMD	×	45.1±1.2	27.7±2.9	2.8±1.3
ATDOC	×	47.3±1.4	31.8±1.4	3.1±0.9
STAR	×	49.8±5.6	24.4±2.4	2.2±2.7
CORAL	×	50.9±2.9	31.7±1.9	1.7±0.4
MCD	×	59.0±3.1	30.7±2.5	1.9±1.9
DRST	×	37.1±6.0	29.5±4.6	1.0±0.7
ICON	×	54.2±1.4	31.1±2.7	4.4±3.2
GroupDRO	✓	61.4±2.7	63.0±2.6	3.4±1.6
GroupDRO (with Tgt)	✓	90.6±0.2	89.3±1.3	73.1±0.3
PDE	✓	57.1±6.6	55.0±5.5	1.3±1.2
Ours (ERM)	×	**87.3**±2.1	**85.0**±4.1	**7.5**±0.5

Table 2: Comparison of test accuracies across Waterbirds, CelebA, and CMNIST. ERM (Src-only) and GroupDRO are trained solely on source domain data, without using any target domain samples. GroupDRO (with Tgt) denotes the setting where both source labeled data and target labeled data are available during training. Boldface indicates the best performance except for models that use labeled target data.

It is well-known that classical UDA methods such as DANN, CDAN, and MK-MMD suffer from poor generalization when spurious correlations exist across domains, as they mainly match marginal input distributions without preserving semantic invariance (Johansson et al., 2019; Zhao et al., 2019). This often aligns spurious attributes (e.g., background, gender, or color) instead of causally related features, which can hurt generalization to the target domain. Pseudo-labeling methods such as STAR and ATDOC are also vulnerable, since spurious correlations can induce inaccurate pseudo-labels that propagate errors during training.

In contrast, our method does not rely on reducing the distance between marginal input distributions. Instead, it explicitly considers a set of plausible target distributions through a distributional ambiguity set and optimizes for the worst-case target risk within this set. By accounting for distributional variations that may arise from spurious attributes, this formulation guides the model to focus on predictive features that are invariant across the plausible target distributions, thereby enhancing robustness to spurious correlations. As a result, it significantly outperforms existing baselines, particularly in scenarios where spurious correlations dominate and unlabeled target data is extremely limited. Compared to ERM, it improves target domain accuracy by $+38.9\%$ on Waterbirds, $+49.5\%$ on CelebA, and $+6.6\%$ on CMNIST.

5 Conclusion

We presented a DRO framework for UDA that jointly models uncertainty in both the target covariate distribution and the conditional label distribution. By formulating an ambiguity set over mixtures of source conditionals and allowing controlled perturbations in target inputs, our method directly addresses two challenges that have gained increasing attention in recent research: the scarcity of unlabeled target data and the presence of spurious correlations in the source domain. Extensive experiments on digit recognition and spurious correlation benchmarks demonstrated consistent and substantial performance gains over strong baselines, highlighting the method's robustness and effectiveness. Our method provides a principled and computationally efficient solution for building models that remain reliable under severe distribution shifts.

ACKNOWLEDGMENTS

REFERENCES

M Amini and Patrick Gallinari. Semi-supervised learning with an explicit label-error model for misclassified data. In *Proc. International Joint Conferences on Artificial Intelligence*, pp. 555–560, 2003.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *ArXiv:1907.02893*, 2019.

Akram S Awad and George K Atia. Distributionally robust domain adaptation. In *Proc. International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2023.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proc. European Conference on Computer Vision*, pp. 456–473, 2018.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Jose Blanchet, Karthyek Murthy, and Viet Anh Nguyen. Statistical analysis of Wasserstein distributionally robust estimators. In *Tutorials in Operations Research: Emerging optimization methods and modeling techniques with applications*, pp. 227–254, 2021.

Tiffany Tianhui Cai, Hongseok Namkoong, and Steve Yadlowsky. Diagnosing model performance under distribution shift. *ArXiv:2303.02011*, 2023.

Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proc. Computer Vision and Pattern Recognition*, pp. 7354–7362, 2019.

Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift. In *Proc. Advances in Neural Information Processing Systems*, volume 33, pp. 21061–21071, 2020.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *ArXiv:1705.08848*, 2017a.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017b.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proc. European Conference on Computer Vision*, pp. 447–463, 2018.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 72(3):1177–1191, 2024.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *ArXiv:2007.01434*, 2020.

Zijian Guo. Statistical inference for maximin effects: Identifying stable associations across multiple studies. *Journal of the American Statistical Association*, 119(547):1968–1984, 2024.

Zijian Guo, Zhenyu Wang, Yifan Hu, and Francis Bach. Statistical inference for conditional group distributionally robust optimization with cross-entropy loss. *ArXiv:2507.09905*, 2025.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *ArXiv:2004.10964*, 2020.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Proc. International Conference on Machine Learning*, pp. 1989–1998, 2018.

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 2002.

Xin Jin, Hongyu Zhu, Siyuan Li, Zedong Wang, Zicheng Liu, Juanxi Tian, Chang Yu, Huafeng Qin, and Stan Z Li. A survey on mixup augmentations and beyond. *ArXiv:2409.05202*, 2024.

Ying Jin, Kevin Guo, and Dominik Rothenhäusler. Diagnosing the role of observable distribution shift in scientific replications. *ArXiv:2309.01056*, 2023.

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *Proc. Artificial Intelligence and Statistics*, 2019.

Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *Proc. International Conference on Information Processing in Medical Imaging*, pp. 597–609. Springer, 2017.

Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *Proc. International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *Proc. International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *Proc. International Conference on Machine Learning*, pp. 5468–5479, 2020.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. In *Proc. Advances in Neural Information Processing Systems*, volume 33, pp. 728–740, 2020.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*, pp. 896, 2013.

Zijian Li, Ruichu Cai, Guangyi Chen, Boyang Sun, Zhifeng Hao, and Kun Zhang. Subspace identification for multi-source domain adaptation. In *Proc. Advances in Neural Information Processing Systems*, volume 36, pp. 34504–34518, 2023.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proc. International Conference on Machine Learning*, pp. 6028–6039, 2020.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proc. International Conference on Machine Learning*, pp. 6781–6792, 2021.

Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Latent covariate shift: Unlocking partial identifiability for multi-source domain adaptation. *Transactions on Machine Learning Research*, 2025.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. International Conference on Computer Vision*, pp. 3730–3738, 2015.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *Proc. International Conference on Machine Learning*, pp. 97–105, 2015.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Proc. Advances in Neural Information Processing Systems*, volume 31, pp. 1640–1650, 2018.

Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *Proc. Computer Vision and Pattern Recognition*, pp. 9111–9120, 2020.

Nicolai Meinshausen and Peter Bühlmann. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In *Proc. Advances in Neural Information Processing Systems*, volume 33, pp. 20673–20684, 2020.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Proc. Advances in Neural Information Processing Systems*, volume 29, pp. 2208–2216, 2016.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, pp. 7, 2011.

Duo Peng, Qiuhong Ke, ArulMurugan Ambikapathi, Yasin Yazici, Yinjie Lei, and Jun Liu. Unsupervised domain adaptation via domain-adaptive diffusion. *IEEE Transactions on Image Processing*, 2024.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *ArXiv:1911.08731*, 2019.

Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, et al. Extending the wilds benchmark for unsupervised adaptation. *ArXiv:2112.05090*, 2021.

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proc. Computer Vision and Pattern Recognition*, pp. 3723–3732, 2018.

Yuge Shi, Jeffrey Seely, Philip HS Torr, Narayanaswamy Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *ArXiv:2104.09937*, 2021.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A Dirt-t approach to unsupervised domain adaptation. *ArXiv:1802.08735*, 2018.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Proc. Advances in Neural Information Processing Systems*, volume 33, pp. 19339–19352, 2020.

Matthew Staib and Stefanie Jegelka. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS Workshop on Machine Learning and Computer Security*, volume 3, pp. 4, 2017.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Proc. European Conference on Computer Vision*, pp. 443–450, 2016.

Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. Domain adaptation with conditional distribution matching and generalized label shift. In *Proc. Advances in Neural Information Processing Systems*, volume 33, pp. 19276–19289, 2020.

Lianghao Tan, Zhuo Peng, Yongjia Song, Xiaoyi Liu, Huangqi Jiang, Shubing Liu, Weixi Wu, and Zhiyuan Xiang. Unsupervised domain adaptation method based on relative entropy regularization and measure propagation. *Entropy*, 27(4):426, 2025.

Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proc. Computer Vision and Pattern Recognition*, pp. 8725–8735, 2020.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proc. Computer Vision and Pattern Recognition*, pp. 7167–7176, 2017.

Georg Volk, Stefan Müller, Alexander Von Bernuth, Dennis Hospach, and Oliver Bringmann. Towards robust cnn-based object detection through augmentation with synthetic rain variations. In *Proc. IEEE Intelligent Transportation Systems Conference*, pp. 285–292. IEEE, 2019.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Zhenyu Wang, Peter Bühlmann, and Zijian Guo. Distributionally robust machine learning with multi-source data. *The Annals of Statistics*, 2025.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. *ArXiv:2010.03622*, 2020.

Xing Wei, Wenhao Jiang, Fan Yang, Chong Zhao, Yang Lu, Benhong Zhang, and Xiang Bi. Unsupervised domain adaptation via causal-contrastive learning. *The Journal of Supercomputing*, 81 (5):719, 2025.

Keru Wu, Yuansi Chen, Wooseok Ha, and Bin Yu. Prominent roles of conditionally invariant components in domain adaptation: Theory and algorithms. *Journal of Machine Learning Research*, 26(110):1–92, 2025.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proc. Computer Vision and Pattern Recognition*, pp. 10687–10698, 2020.

Xinyue Xu, Yueying Hu, Hui Tang, Yi Qin, Lu Mi, Hao Wang, and Xiaomeng Li. Concept-based unsupervised domain adaptation. In *Proc. International Conference on Machine Learning*, 2025.

Jiawen Yang, Shuhao Chen, Yucong Duan, Ke Tang, and Yu Zhang. Heterogeneous-modal unsupervised domain adaptation via latent space bridging. *ArXiv:2506.15971*, 2025.

Shiqi Yang, Joost Van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. In *Proc. Advances in Neural Information Processing Systems*, volume 34, pp. 29393–29405, 2021.

Zhongqi Yue, Qianru Sun, and Hanwang Zhang. Make the u in uda matter: Invariant consistency learning for unsupervised domain adaptation. In *Proc. Advances in Neural Information Processing Systems*, volume 36, pp. 26991–27004, 2023.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proc. European Conference on Computer Vision*, pp. 818–833, 2014.

Keyao Zhan, Xin Xiong, Zijian Guo, Tianxi Cai, and Molei Liu. Domain adaptation optimized for robustness in mixture populations. *ArXiv:2407.20073*, 2024.

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *ArXiv:2203.01517*, 2022.

Yue Zhang, Mingyue Bin, Yuyang Zhang, Zhongyuan Wang, Zhen Han, and Chao Liang. Link-based contrastive learning for one-shot unsupervised domain adaptation. In *Proc. Computer Vision and Pattern Recognition*, pp. 4916–4926, 2025.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *Proc. International Conference on Machine Learning*, 2019.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. In *Proc. Pattern Analysis and Machine Intelligence*, volume 40, pp. 1452–1464, 2017.

A Appendix

A.1 Rationale behind Proposition 3.1

In this subsection, we provide the detailed rationale of the tractable reformulation stated in Proposition 3.1. With the ambiguity set $\mathcal{Q}$ defined in (2), the inner maximization in the DRO objective can be expressed as

$$\sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[\ell\left(f^\theta(X),Y\right)] = \sup_{Q\in\mathcal{Q}} \mathbb{E}_{Q_X}\mathbb{E}_{Q_{Y|X}}[\ell\left(f^\theta(X),Y\right)] \tag{8}$$

$$= \sup_{\substack{\beta\in\Delta_{K-1},\\ \|\beta-\bar{\beta}\|_2\le\epsilon_2}} \sup_{W_\infty(Q_X,\hat{P}_X^{\mathrm{tg}})\le\epsilon_1} \mathbb{E}_{Q_X}\mathbb{E}_{Q_{Y|X}^\beta}\left[\ell\left(f^\theta(X),Y\right)\right], \tag{9}$$

where

$$Q^{\beta}_{Y|X} = \sum_{k=1}^{K} \beta_k \hat{P}^{(k)}_{Y|X}.$$

We recall the following lemma from Staib & Jegelka (2017), which will be useful in our reformulation.

Lemma A.1.1 (*Staib & Jegelka (2017), Proposition 3.1*)**.** *Suppose that the cost function $c(\cdot,\cdot)$ used to define the Wasserstein distance is a metric on $\mathcal{X}$. Let P be a Borel probability measure on $\mathcal{X}$, and let $f : \mathcal{X} \to \mathbb{R}$ be a measurable function. Then, for any $\epsilon \geq 0$,*

$$\sup_{W_\infty(Q,P)\leq\epsilon} \mathbb{E}_Q\left[f(X)\right] = \mathbb{E}_P\left[\sup_{x'\in B_\epsilon(X)} f(x')\right],$$

where $B_\epsilon(x) = \{x' \in \mathcal{X} : c(x,x') \leq \epsilon\}$ and $\mathbb{E}_P$ denotes the expectation under $X \sim P$.

By Lemma A.1.1, we have

$$\begin{aligned}
&\sup_{W_\infty(Q_X,\hat{P}^{\mathrm{tg}}_X)\leq\epsilon_1} \mathbb{E}_{Q_X}\left[\mathbb{E}_{Q^{\beta}_{Y|X}}\left[\ell\left(f^{\theta}(X),Y\right)\right]\right] \\
&\quad= \mathbb{E}_{\hat{P}^{\mathrm{tg}}_X}\left[\sup_{x:\|z(x)-z(X)\|_2\leq\epsilon_1} \mathbb{E}_{Q^{\beta}_{Y|X}}\left[\ell\left(f^{\theta}(X),Y\right)\right]\right] \\
&\quad= \mathbb{E}_{\hat{P}^{\mathrm{tg}}_X}\left[\sup_{x:\|z(x)-z(X)\|_2\leq\epsilon_1} \mathbb{E}_{Q^{\beta}_{Y|X}}\left[\ell\left(f^{\theta}_Z(z(X)),Y\right)\right]\right] \\
&\quad\leq \mathbb{E}_{\hat{P}^{\mathrm{tg}}_X}\left[\sup_{z':\|z'-z(X)\|_2\leq\epsilon_1} \mathbb{E}_{Q^{\beta}_{Y|X}}\left[\ell\left(f^{\theta}_Z(z'),Y\right)\right]\right].
\end{aligned}$$

Therefore, (8) is upper-bounded by

$$\sup_{\substack{\beta\in\Delta_{K-1},\\ \|\beta-\bar{\beta}\|_2\leq\epsilon_2}} \mathbb{E}_{\hat{P}^{\mathrm{tg}}_X}\left[\sup_{\|z'-z(X)\|_2\leq\epsilon_1} \mathbb{E}_{Q^{\beta}_{Y|X}}\left[\ell\left(f^{\theta}_Z(z'),Y\right)\right]\right].$$

By treating the probability mass function $\hat{p}^{(k)}_{Y|X}(\cdot \mid X)$ as a probability vector, we define the soft pseudo-label vector as

$$y^{\circ}(\beta,x) := \sum_{k=1}^{K} \beta_k\, \hat{p}^{(k)}_{Y|X}(\cdot \mid x). \tag{10}$$

Recall that the cross-entropy loss is defined by

$$\ell\left(f^{\theta}_Z(z'),Y\right) = -\sum_{y\in\mathcal{Y}} \mathbb{I}[Y=y]\log \pi^{\theta}_Z(y \mid z'),$$

where $\pi^{\theta}_Z(\cdot \mid z')$ denotes the predicted class probability vector obtained by applying the softmax function to $f^{\theta}_Z(z')$ and $\mathbb{I}[\cdot]$ is the indicator function. Therefore, we have

$$\begin{aligned}
\mathbb{E}_{Q^{\beta}_{Y|X=x}}\left[\ell\left(f^{\theta}_Z(z'),Y\right)\right] &= \mathbb{E}_{Q^{\beta}_{Y|X=x}}\left[-\sum_{y\in\mathcal{Y}} \mathbb{I}[Y=y]\log \pi^{\theta}_Z(y \mid z')\right] \\
&= -\sum_{y\in\mathcal{Y}}\left(\sum_{k=1}^{K} \beta_k\, \hat{p}^{(k)}_{Y|X}(y \mid x)\right)\log \pi^{\theta}_Z(y \mid z') \\
&= \ell\left(f^{\theta}_Z(z'), y^{\circ}(\beta,x)\right).
\end{aligned} \tag{11}$$

Here, we slightly abuse the notation for $\ell(\cdot,\cdot)$, interpreting $\ell(f^{\theta}_Z(z'), y^{\circ}(\beta,x))$ as the cross-entropy between the two probability vectors $\pi^{\theta}_Z(\cdot \mid z')$ and $y^{\circ}(\beta,x)$.

Combining the above, we obtain the following surrogate objective for (8):

$$\sup_{\substack{\beta\in\Delta_{K-1},\\ \|\beta-\bar{\beta}\|_2\le\epsilon_2}} \mathbb{E}_{\hat{P}_X^{\mathrm{tg}}}\left[\sup_{\|z'-z(X)\|_2\le\epsilon_1} \ell\left(f_Z^\theta(z'), y^\circ(\beta, X)\right)\right]. \tag{12}$$

A.2 Exponentiated gradient ascent for updating β

We first review the exponentiated gradient method briefly and then describe its application to our β-update procedure.

Exponentiated gradient ascent method The exponentiated gradient ascent method can be interpreted as a mirror ascent step over the simplex, obtained by maximizing the first-order approximation of the objective with a KL divergence regularization. Specifically, let $\mathcal{L}(\beta)$ be a differentiable objective function. Then, the exponentiated gradient ascent update from $\beta^{(t)}$ is given by

$$\beta^{(t+1)} = \arg\max_{\beta\in\Delta_{K-1}} \left\{\mathcal{L}(\beta^{(t)}) + \nabla_\beta\mathcal{L}(\beta^{(t)})^\top(\beta-\beta^{(t)}) - \frac{1}{\eta_\beta} D_{\mathrm{KL}}\Big(\beta \,\|\, \beta^{(t)}\Big)\right\},$$

where ∇_β denotes the gradient of the mapping $\beta \mapsto \mathcal{L}(\beta)$, and $\eta_\beta > 0$ is the step size. Since the terms involving $\beta^{(t)}$ are constant with respect to β, this is equivalent to

$$\beta^{(t+1)} = \arg\max_{\beta\in\Delta_{K-1}} \left\{\nabla_\beta\mathcal{L}(\beta^{(t)})^\top\beta - \frac{1}{\eta_\beta} D_{\mathrm{KL}}\Big(\beta \,\|\, \beta^{(t)}\Big)\right\}.$$

By the KKT optimality conditions, the maximization above admits the following closed-form solution:

$$\beta_k^{(t+1)} = \frac{\beta_k^{(t)} \exp\Big(\eta_\beta\left[\nabla_\beta\mathcal{L}(\beta^{(t)})\right]_k\Big)}{\sum_{j=1}^K \beta_j^{(t)} \exp\Big(\eta_\beta\left[\nabla_\beta\mathcal{L}(\beta^{(t)})\right]_j\Big)}, \qquad k=1,\dots,K, \tag{13}$$

where $[A]_i$ denotes i-th component of A. See Kivinen & Warmuth (1997) for details.

Detailed procedure for updating β We now revisit our optimization problem in (3). Since the constraint $\|\beta-\bar{\beta}\|_2 \le \epsilon_2$ is handled by projection after the ascent step, we focus only on updating β over the simplex Δ_{K-1}. Suppose that x_i^{tg} and z_i' are given. Then, we need to solve

$$\sup_{\beta\in\Delta_{K-1}} \left[\ell\big(f_Z^\theta(z_i'),\, y^\circ(\beta, x_i^{\mathrm{tg}})\big)\right].$$

This problem can be expressed as:

$$\begin{aligned}
\sup_{\beta\in\Delta_{K-1}} \ell\big(f_Z^\theta(z_i'),\, y^\circ(\beta, x_i^{\mathrm{tg}})\big) &= \sup_{\beta\in\Delta_{K-1}} \mathbb{E}_{Q^\beta_{Y|X=x_i^{\mathrm{tg}}}}\left[\ell\big(f_Z^\theta(z_i'), Y\big)\right] \\
&= \sup_{\beta\in\Delta_{K-1}} \sum_{k=1}^K \beta_k\, \mathbb{E}_{\hat{P}^{(k)}_{Y|X=x_i^{\mathrm{tg}}}}\left[\ell\big(f_Z^\theta(z_i'), Y\big)\right] \\
&= \sup_{\beta\in\Delta_{K-1}} \sum_{k=1}^K \beta_k\, \ell\Big(f_Z^\theta(z_i'),\, \hat{p}^{(k)}_{Y|X}(\cdot \mid x_i^{\mathrm{tg}})\Big),
\end{aligned}$$

where this expression follows from (4) and (11).

Setting

$$\mathcal{L}(\beta) := \sum_{k=1}^K \beta_k\, \ell\Big(f_Z^\theta(z_i'),\, \hat{p}^{(k)}_{Y|X}(\cdot \mid x_i^{\mathrm{tg}})\Big),$$

and applying the closed-form solution in (13), we obtain the following update rule:

$$\beta_k \leftarrow \frac{\beta_k \exp\Big(\eta_\beta\, \ell\Big(f_Z^\theta(z_i'),\, \hat{p}^{(k)}_{Y|X}(\cdot \mid x_i^{\mathrm{tg}})\Big)\Big)}{\sum_{j=1}^K \beta_j \exp\Big(\eta_\beta\, \ell\Big(f_Z^\theta(z_i'),\, \hat{p}^{(j)}_{Y|X}(\cdot \mid x_i^{\mathrm{tg}})\Big)\Big)}, \quad k=1,\dots,K.$$

A.3 EXPERIMENTS DETAILS

A.3.1 DIGIT BENCHMARKS

MNIST (LeCun et al., 2002): A widely used benchmark dataset of handwritten digit recognition, consisting of grayscale images of digits from 0 to 9. Each image is of size 28×28 pixels, yielding 784-dimensional input vectors when flattened. The dataset contains a total of 70,000 images, with 60,000 images allocated for training and 10,000 images reserved for testing. The digit classes are balanced across 10 categories, corresponding to the labels 0 through 9. Due to its simplicity and accessibility, MNIST has been extensively used for evaluating classification algorithms, representation learning methods, and as a canonical testbed for new machine learning models.

SVHN (Netzer et al., 2011): The Street View House Numbers dataset, designed for digit recognition in natural scenes. It consists of cropped color images of digits (32×32 pixels) obtained from Google Street View house numbers. SVHN provides 73,257 training images, 26,032 test images, and an additional set of 531,131 extra training images to facilitate large-scale training. The dataset covers 10 digit classes (0–9) and presents greater variability than MNIST due to cluttered backgrounds, varying illumination, and diverse font styles, making it a challenging and widely used benchmark in digit classification and domain adaptation studies.

USPS (Hull, 2002): The U.S. Postal Service (USPS) dataset is a benchmark collection of handwritten digits originally scanned from postal envelopes. Each image is a grayscale 16×16 pixel representation of digits ranging from 0 to 9, resulting in compact 256-dimensional input vectors when flattened. The dataset provides 7,291 training images and 2,007 test images, spanning 10 classes that are approximately balanced across digits. Compared to MNIST, USPS offers lower-resolution images and displays noticeable stylistic variations in handwriting, including slant, thickness, and shape differences, which make recognition more challenging. Due to these characteristics, USPS is often used alongside MNIST and SVHN as a complementary benchmark in digit classification tasks and as a target or source domain in domain adaptation studies.

(a) MNIST (b) SVHN (c) USPS

Figure 1: Example images from the CMNIST dataset.

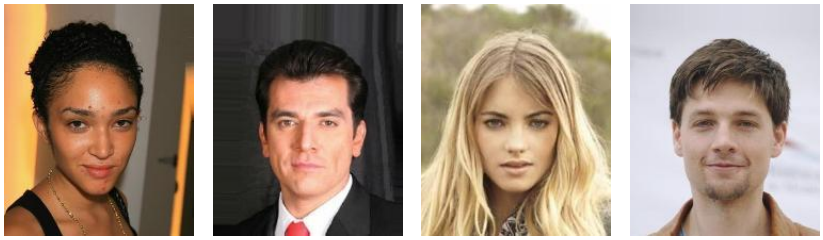
A.3.2 SPURIOUS CORRELATION BENCHMARKS

Waterbirds (Sagawa et al. (2019)): The Waterbirds dataset is designed to study the impact of spurious correlations in image classification. It is constructed by overlaying bird images from the CUB-200-2011 dataset (Wah et al. (2011)) onto background scenes from the Places dataset (Zhou et al. (2017)). This process induces a strong but spurious correlation between bird type and background. Specifically, the dataset is divided into four groups based on bird type (landbird vs. waterbird) and background (land vs. water): $g_1 = \{\text{landbird, land}\}$, $g_2 = \{\text{landbird, water}\}$, $g_3 = \{\text{waterbird, land}\}$, and $g_4 = \{\text{waterbird, water}\}$. Most samples belong to the aligned groups (g_1 and g_4), while the misaligned groups (g_2 and g_3) are relatively rare. This imbalance creates minority groups that highlight the difficulty of learning invariant predictors under spurious correlations. In our experiments, we follow prior work by treating the three majority groups as the single-source domain and the minority group (g_3, waterbirds with a land background) as the target domain. The training set contains 4,739 source images and 56 unlabeled target images, while the target test set includes 642 samples, providing a controlled evaluation of out-of-distribution generalization.

(a) Landbirds, Land (b) Landbirds, Water (c) Waterbirds, Land (d) Waterbirds, Water

Figure 2: Example images from the Waterbirds dataset.

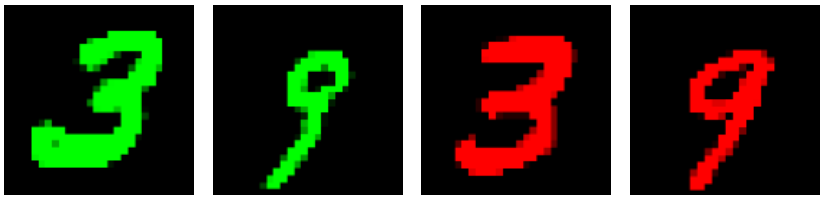
CelebA (Liu et al. (2015)): The CelebA dataset contains large-scale celebrity face images annotated with multiple binary attributes. We consider the task of classifying blond vs. non-blond hair, where gender acts as a spurious attribute correlated with hair color. This naturally induces four groups based on hair color (blond vs. non-blond) and gender (male vs. female): $g_1 = \{$non-blond, female$\}$, $g_2 = \{$non-blond, male$\}$, $g_3 = \{$blond, female$\}$, and $g_4 = \{$blond, male$\}$. Most samples belong to g_1, g_2, and g_3, while the minority group g_4 (blond-haired males) is underrepresented, reflecting the spurious correlation between hair color and gender. The three majority groups are combined to form the single-source domain, and the minority group is treated as the target domain. The dataset provides $161{,}383$ labeled source training images, $1{,}387$ unlabeled target training images, and 642 target test images for evaluation.



(a) Non-blond, Female (b) Non-blond, Male (c) Blond, Female (d) Blond, Male

Figure 3: Example images from the CelebA dataset.

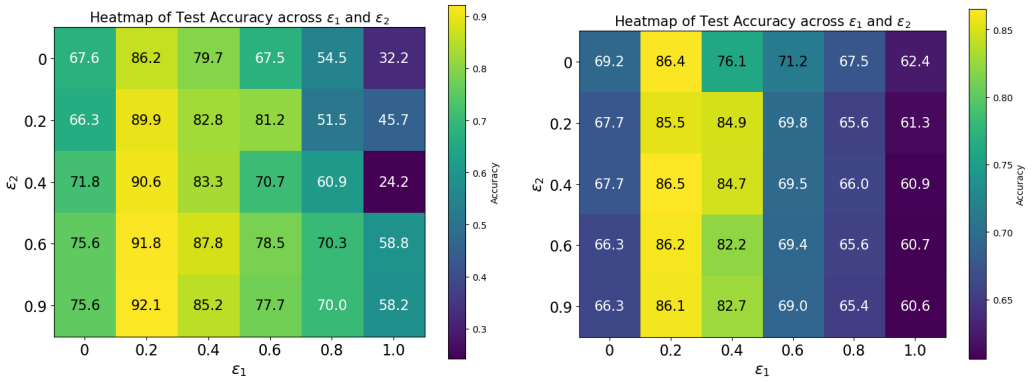
Colored MNIST (CMNIST) (Arjovsky et al. (2019)): The Colored MNIST dataset is a synthetic variant of MNIST designed to study the effect of spurious correlations. We consider a binary classification task using digits 0 and 1, which yields four groups based on digit identity and color: $g_1 = \{$digit 0, red$\}$, $g_2 = \{$digit 0, green$\}$, $g_3 = \{$digit 1, red$\}$, and $g_4 = \{$digit 1, green$\}$. Most samples belong to the aligned groups (g_1, g_4, and g_3), while the minority group g_2 (digit 0 with green color) is underrepresented, capturing the spurious correlation between digit and color. In our experiments, the three majority groups are combined to form the single-source domain, and the minority group is used as the target domain. The dataset provides $26{,}002$ labeled source training images, $2{,}998$ unlabeled target training images, and $8{,}966$ target test images.



(a) 0, Green (b) 1, Green (c) 0, Red (d) 1, Red

Figure 4: Example images from the CMNIST dataset.

A.3.3 IMPACT OF (ϵ_1, ϵ_2) ON MODEL PERFORMANCE



(a) MNIST → USPS (10^2) (b) USPS → MNIST (10)

Figure 5: Heatmaps of average test accuracy across (ϵ_1,ϵ_2) for MNIST → USPS and USPS → MNIST. The numbers 10^2 and 10 denote the unlabeled target samples per class.

In this subsection, we demonstrate how the hyperparameters ϵ_1 and ϵ_2 influence model performance. Figures 5(a) and 5(b) present heatmaps of the average test accuracy under different settings of ϵ_1 and ϵ_2. Figure 5(a) shows results for the MNIST → USPS task with 10^2 unlabeled target samples per class, while Figure 5(b) illustrates the USPS → MNIST task with only 10 unlabeled target samples per class.

The figures show that the adjustment of the hyperparameters ϵ_1 and ϵ_2 significantly increases the average test accuracy compared to the baseline $(\epsilon_1, \epsilon_2) = (0, 0)$. This highlights that consideration for uncertainty in both feature and conditional distributions enhances performance by improving the model's generalization on target domain.

Importantly, the relative impact of the two hyperparameters differs across datasets. As shown in Figure 5(a), both ϵ_1 and ϵ_2 contribute jointly to performance improvement on MNIST → USPS, indicating that robustness is enhanced by allowing uncertainty in both feature and conditional distributions. In contrast, Figure 5(b) highlights that when unlabeled target data N^{tg} is very limited, ϵ_2 becomes particularly critical: controlling covariate uncertainty substantially improves performance by compensating for the scarcity of unlabeled target data.

A.3.4 BASELINE DETAILS

- ERM (Src-only) The standard empirical risk minimization approach that optimizes average accuracy on the source training set. While simple and widely used, ERM does not incorporate any robust objective and cannot exploit unlabeled target data, making it insufficient under domain shift.
- CDAN (Long et al., 2018): Extends DANN by conditioning the domain discriminator on both feature representations and classifier predictions. This coupling captures multimodal structures in the feature space, leading to more effective domain alignment.
- MK-MMD (Long et al., 2015): Minimizes the maximum mean discrepancy (MMD) across multiple kernels between source and target feature distributions. By leveraging multiple kernels, MK-MMD adapts flexibly to complex distribution shifts.
- CORAL (Sun & Saenko, 2016): Matches the second-order statistics (covariances) of source and target features to reduce domain discrepancy. This moment-matching approach is simple, efficient, and widely adopted in UDA.
- MCD (Saito et al., 2018): Employs two classifiers with a shared feature extractor. By maximizing their prediction discrepancy on target samples and then minimizing it, MCD encourages the extractor to learn target-discriminative features.

- ATDOC (Tang et al., 2020): Adapts classifiers by leveraging target-domain pseudo-labels in an adversarial manner. It progressively refines pseudo-labels to improve alignment and robustness in the absence of ground-truth target labels.
- STAR (Lu et al., 2020): Introduces stochastic classifiers to model diverse decision boundaries. By averaging predictions across multiple classifiers, STAR improves stability and robustness under domain shifts.
- Group DRO (Sagawa et al., 2019): A robust optimization method that minimizes the worst-case loss across predefined groups. Since it relies on explicit group labels in the training data, Group DRO cannot leverage unlabeled target data, which limits its applicability in unsupervised domain adaptation.
- Group DRO (with Tgt) (Sagawa et al., 2019): A robust optimization method that minimizes the worst-case loss across predefined groups. In this variant, we assume access to group labels from both the source and target domains during training. This setup is not realistic in unsupervised domain adaptation, since target group labels are unavailable in practice.
- ICON (Yue et al., 2023): Focuses on unsupervised domain adaptation by enforcing invariant consistency across source and target predictions. ICON leverages consistency regularization to learn representations robust to distributional shifts, achieving state-of-the-art performance in UDA.

A.4 Disclosure on LLM Usage

We used large language models solely to aid in polishing the writing of this paper.